SAPIENZA
UNIVERSITÀ DI ROMA

DIPARTIMENTO DI INGEGNERIA INFORMATICA AUTOMATICA E GESTIONALE
ANTONIO RUBERTI

Master of Science in Engineering in Computer Science (MSE-CS)

Seminars in Software and Services for the Information Society

Umberto Nanni

Introduction to Data Mining

Data Mining

- born before the Data Warehouse
- collection of techniques from: Artificial Intelligence, Pattern Recognition, Statistics (e.g., genetic algorithms, fuzzy logic, expert systems, neural networks, etc.)
- targets:
  - descriptive goals: identify patterns of behavior, cause-effect relationships, classifying individuals, etc.
  - predictive goals: predict trends, to classify individuals according to risk, etc.

Some applications for Data Mining

- Data Analysis and Decision Support Systems
- Market Analysis and Marketing
  - Target Marketing, Customer Relationship Management (CRM), Market Basket Analysis (MBA), market segmentation
- Analysis and risk management
  - reliability forecasts, user loyalty, quality control, ...
  - detection of frauds and unusual patterns (outliers)
- Text Mining
- Web Mining, ClickStream Analysis
- Genetic engineering, DNA interpretation, ...

Data Mining: associative rules

IF X ("the customer purchases beer")
THEN Y ("the customer purchases diaper")

$$X \rightarrow Y$$

Support (what fraction of individual follows the rule):

$$s = \frac{|X \cap Y|}{|all|}$$

$s(X \rightarrow Y) = F(X \wedge Y)$

Confidence (what fraction of individual to whom the rule applies, follows the rule):

$$c = \frac{|X \cap Y|}{|X|}$$

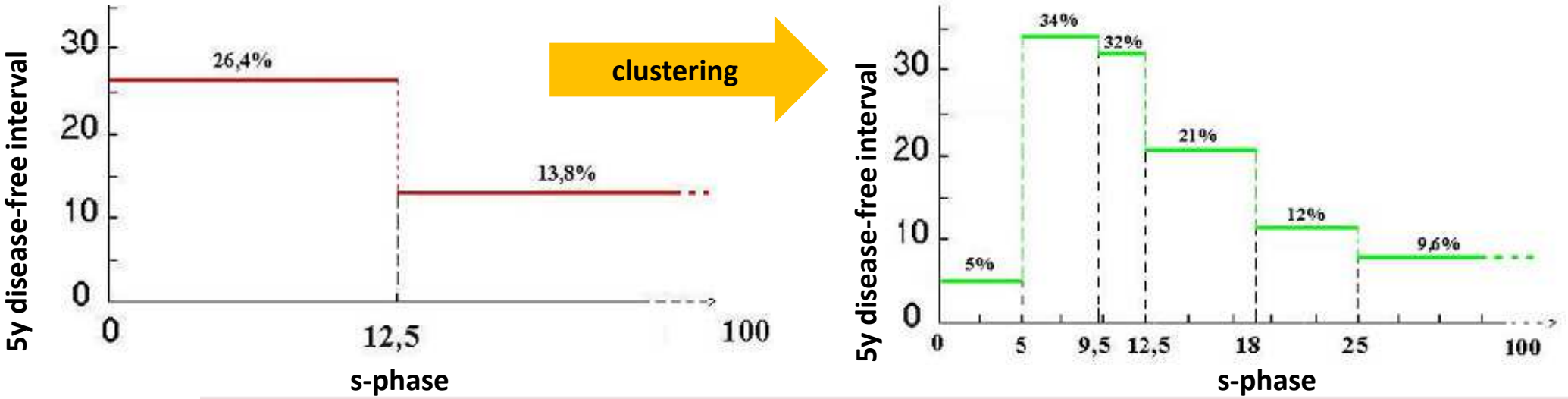
$c(X \rightarrow Y) = F(Y \mid X)$

Range: economics (e.g.: *market basket analysis*), telecommunication, health care, …

Data Mining: clustering

- identify similarities, spot heterogeneity in the distribution in order to define homogeneous groups (unsupervised learning)
- search clusters based on
  - distribution of population
  - a notion of “distance”

Example: DFI – Disease-Free Interval (5 years)
(collaboration with Ist. Regina Elena, Roma)



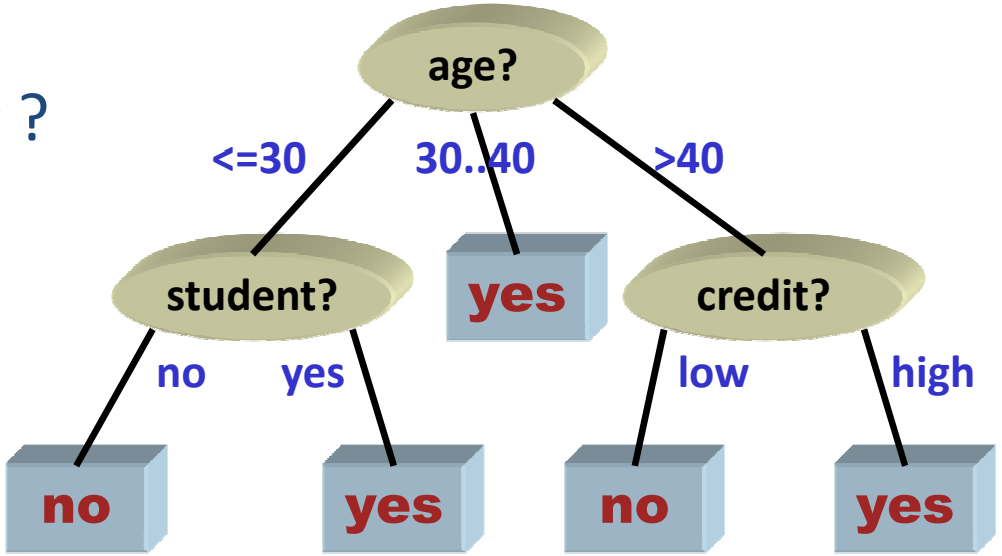
Data Mining: decision tree

Determine the causes of an interesting phenomenon (with a set of output values), sorted by relevance

- internal node: attribute value to be appraised
- branching: value (or value interval) for an attribute
- leave: one of the possible output values

Example:

will the customer buy a computer ?



Data Mining: time sequences

- spot recurrent / unusual patterns in time sequences
- feature prediction

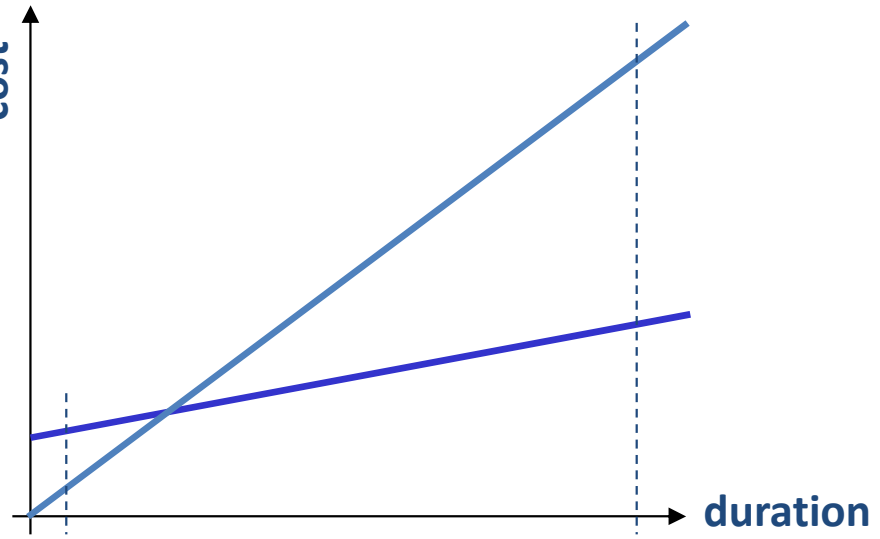
Example (Least Cost Routing): routing a telephone call over the cheapest available connection
(coooperation with Between – consulting firm)

KEY QUESTION:
given an outbound call from an internal line X toward an external number Y, how long the call?

Rates:

connection fee

flat rate



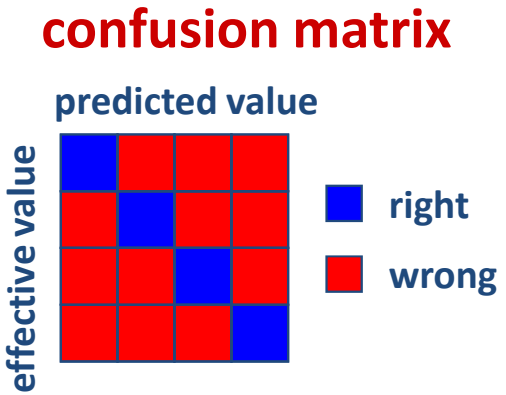
Neural Networks

Problem:

can you write a program which recognizes human writing of capital letters...

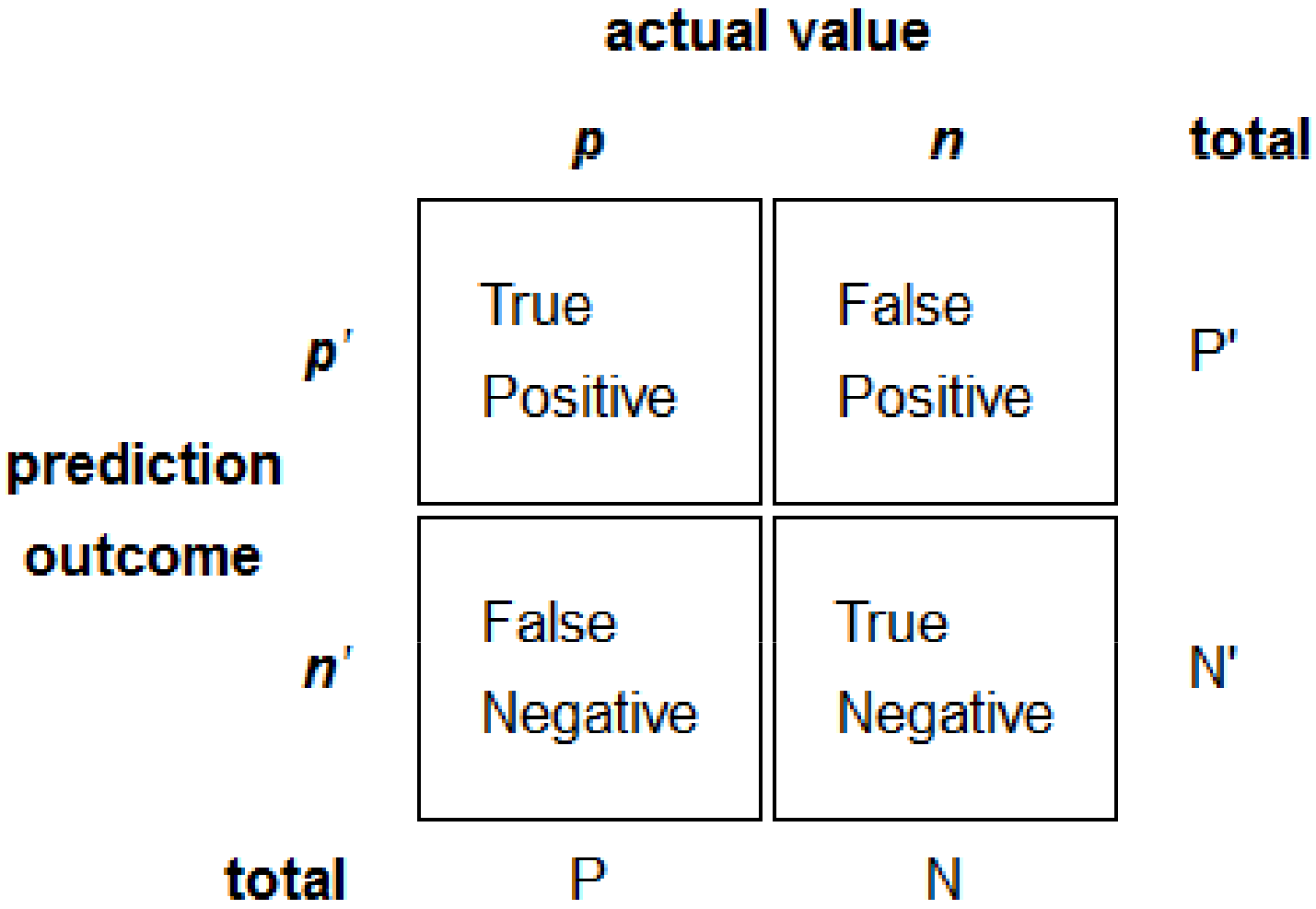
?

Data Mining: “interesting” results



- Simplicity - For example:
  - length of rules (associative)
  - size (decision tree)
- Certainty - For example:
  - confidence (Association Rules): $c(X \rightarrow Y) = \#(X \text{ and } Y) / \#(X)$
  - reliability of classification
- Usefulness - For example:
  - support (Association Rules) $s(X \rightarrow Y) = \#(X \text{ and } Y) / \#(ALL)$
- Novelty - For example:
  - not known previously
  - surprising
  - subsumption of other rules (included as special cases)

Confusion matrix



Confusion matrix & Terminology

Positive (P), Negative (N)

True Positive (TP), True Negative (TN)

False Positive (FP), False Negative (FN)

True Positive Rate [**sensitivity**, **recall**] $TPR = TP / P = TP / (TP+FN)$

False Positive Rate $FPR = FP / N = FP / (FP + TN)$

ACCuracy $ACC = (TP + TN) / (P + N)$

SPeCificity (True Negative Rate) $SPC = TN / N = TN / (FP + TN) = 1 - FPR$

Positive Predictive Value [**precision**] $PPV = TP / (TP + FP)$

Negative Predictive Value $NPV = TN / (TN + FN)$

False Discovery Rate $FDR = FP / (FP + TP)$

ROC curve

Receiver Operating Characteristic

(from signal detection theory)

Fundamental tool for evaluation of a learning algorithm.

Y axis: True Positive Rate

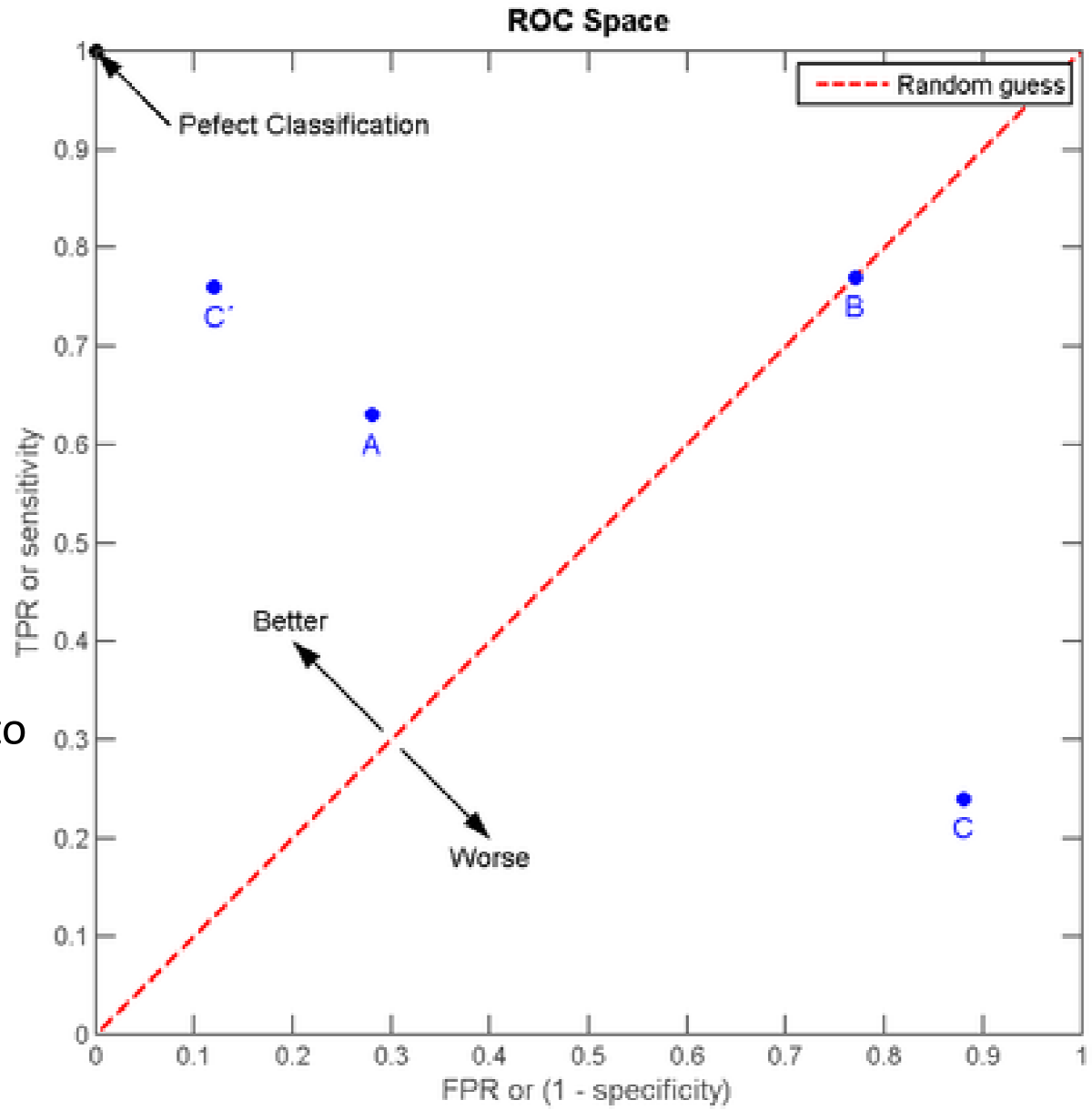
(Sensitivity)

X axis: False Positive Rate

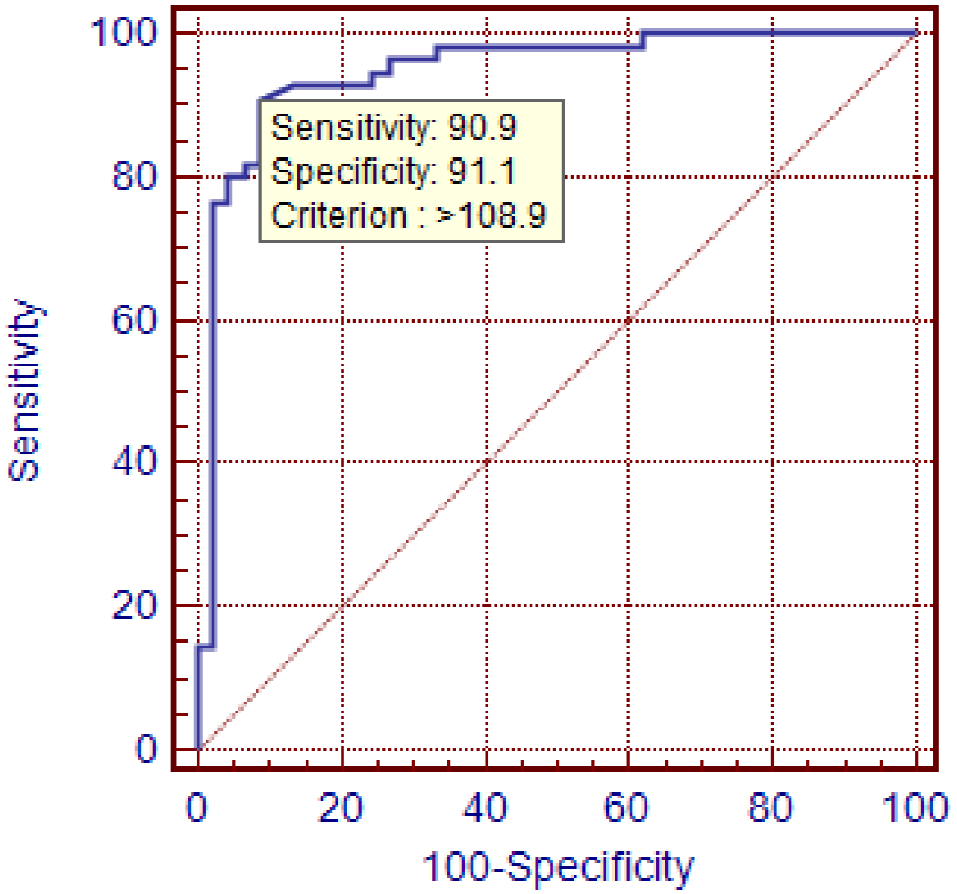
(100-Specificity)

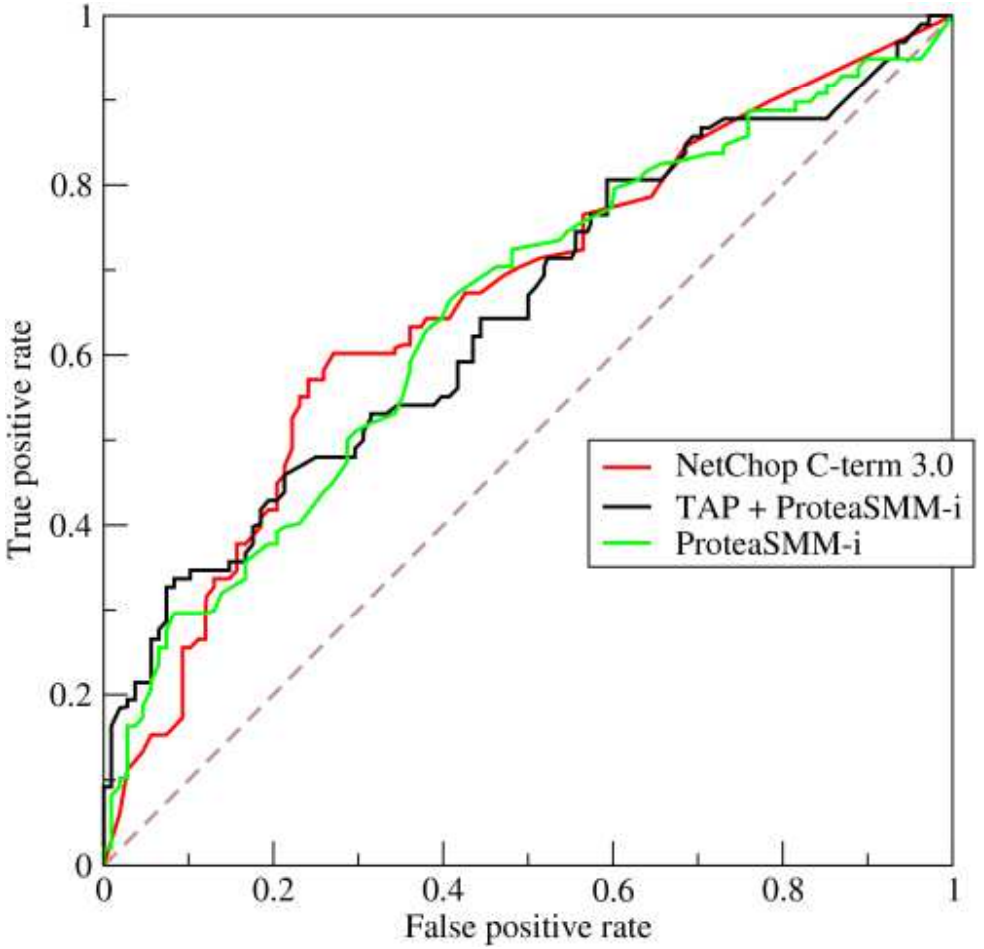
Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

The Area Under the ROC Curve (AUC) is a measure of how well a parameter can distinguish between two groups (YES/NO decision).



ROC curve: examples





Mining Rules from Databases – Algorithm: APRIORI

Rakesh Agrawal, Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. *20th International Conference on Very Large Data Bases (VLDB)*, pp.487-499, Santiago, Chile, September 1994.

APRIORI Algorithm:

1. L_1 = { large 1-itemsets }
2. **for** (k = 2; $L_{k-1} \neq \emptyset$; k++) **do begin**
3. C_k = apriori-generate (L_{k-1}) *// Candidates (extending prev. tuples)* — generation
4. **forall** transactions $t \in \mathcal{D}$ **do begin**
5. C_t = subset(C_k , t) *// Candidates contained in t*
6. **forall** candidates $c \in C_t$ **do**
7. c.count++
8. **end**
9. $L_k = \{ c \in C_k \mid c.count \geq minsupport \}$ — pruning
10. **end**
11. ANSWER = $\bigcup_k L_k$